CHAPTER 21

LONGITUDINAL METHODS

Siek-Toon Khoo, Stephen G. West, Wei Wu, and Oi-Man Kwok

The previous chapters in this volume have focused on the measurement of participants using multiple methods, multiple measures, and in multiple situations. In this chapter the focus shifts to the measurement of the same set of participants on multiple occasions, ideally using the same (or equivalent) measurement instruments. This focus on multiple occasions does not fundamentally alter the application of basic concepts and approaches presented in previous chapters (see Eid, chap. 16, this volume; Eid & Diener, chap. 1, this volume). What is new in this chapter is that longitudinal designs explicitly determine the temporal ordering of the observations. This temporal ordering of observations provides an enhanced ability to elucidate stability and change in individuals over time, to study time-related processes, and to establish the direction of hypothesized causal relationships (Dwyer, 1983; Singer & Willett, 2002).

Longitudinal studies are becoming increasingly prominent in several areas of psychology including clinical, community, developmental, personality, and health. For example, Biesanz, West, and Kwok (2003) found that 24% of the studies published in the 2000 and 2001 volumes of the *Journal of Personality: Personality Process and Individual Differences* section and the *Journal of Personality* included two or more waves of data collection. In the area of psychology most focused on issues of stability and change, we found that 32% of the articles in *Developmental Psychology* in 2002 met these minimum criteria for a longitudinal study of two waves of data collection. This compares to only 15% of the articles published in 1990.

A more in-depth review focused on the longitudinal studies in the 2002 volume of *Developmental Psychology* provides a glimpse of current practice (see also Morris, Robinson, & Eisenberg, chap. 25, this volume). The duration of studies ranged from 12 weeks to 28 years. Approximately 25% of the studies collected only two waves of data, whereas approximately 25% of the studies reported 6 or more waves of data collection, with one study collecting more than 50 waves of data. Measures included standardized measures of ability and intelligence; self-, peer, parent, and teacher reports; ratings and counts of behaviors by trained observers; peer nominations; and physical measures such as weight and heart rate. Although most of the studies included a substantial core set of measures that were administered at each wave, some studies used different measures at each measurement wave, precluding the examination of change over time. The majority of articles reported traditional correlation/regression analyses or analysis of variance. Collins and Sayer (2001), McArdle and Nesselroade (2003), and Singer and Willett (2002) have highlighted the potential advantages of newer approaches to the analysis of longitudinal data, yet approaches such as structural equation modeling (approximately 10%) and growth modeling and

We thank Jeremy Biesanz, Patrick Curran, coeditor Michael Eid, Paras Mehta, Roger Millsap, Steven Reise, and an anonymous reviewer for their comments on an earlier version of this chapter.

examination of growth trajectories (approximately 15%) continue to represent a distinct minority of longitudinal studies.

This chapter considers a number of unique issues that arise when measurements are taken on multiple occasions. We begin with a consideration of some desiderata of measurement from cross-sectional research and consider how they may apply in longitudinal research. We then consider three different longitudinal models: (a) autoregressive models that focus on the stability of participants' relative standing on a construct over time; (b) latent trait–state models that partition the variance in measured constructs into relatively stable (trait) and measurement occasion specific (state) components; and (c) growth curve models that estimate individual growth trajectories. Finally, we consider these longitudinal models in light of measurement concerns and indicate some methods through which these concerns can be addressed.

SOME DESIDERATA FOR GOOD MEASUREMENT: LESSONS FROM CROSS-SECTIONAL RESEARCH

Sources on traditional and modern approaches to measurement (Crocker & Algina, 1986; Embretson & Reise, 2000; Lord & Novick, 1968; McDonald, 1999; West & Finch, 1997) have emphasized issues that arise in narrow windows of time that characterize cross-sectional and short-term (test–retest) studies. These approaches have developed several desiderata for good measurement; three are presented following. We also begin to consider how these desiderata may need to be extended for longitudinal studies. In this section we will use the framework of classical test theory and assume that measures have been collected on a numerical scale.

Reliability

In classical test theory the observed score on a measure (Y) can be partitioned into two parts: true score (T) and error (e). In symbols, this is expressed as $Y = T + e$. T can be defined as the mean of a very large number of independent measurements. e is assumed to be random and independent of the value of the true score. The reliability coefficient represents the proportion of the variance in the observed Y scores (σ^2_Y) that is true score variance (σ^2_T),

$$\rho_{YY} = \frac{\sigma_T^2}{\sigma_Y^2}$$

Reliability is an index of the dependability of the measurement. Two measures of reliability are currently widely reported in the literature, coefficient alpha and the test–retest correlation.

Coefficient alpha. When the data are collected on a single measurement occasion, Cronbach's (1951) coefficient alpha (α) is typically reported. Conceptually, α can be thought of as the correlation between two equivalent scales of the same length given at the same time.

Coefficient alpha has several little-known properties that may limit its usefulness in application (Cortina, 1993; Feldt & Brennan, 1989; Schmitt, 1996). First, α assumes that all items are equally good measures of the underlying construct, a condition known as essential tau equivalence (see section on homogeneity for a fuller description). If some items should ideally be weighted more heavily in estimating the true score, then α will underestimate the reliability. Second, α is dependent on test length. For example, if a 10-item scale had an α = .70 and another exactly parallel set of 10 items could be identified, then α for the 20-item scale would be .82. Third, α addresses sources of error that result from the sampling of equivalent items and potential variability *within* the measurement period (e.g., within-test variability in level of concentration). It does not address error resulting from sources that may vary over measurement occasions (e.g., $\rho_{y_i y_{i'}}$, daily changes in mood). Fourth, a high level of α does not indicate that a single dimension has been measured. For example, Cortina showed that if two *orthogonal* dimensions underlie a set of items, even if the intercorrelations between items within each dimension are modest (e.g., = .30), α will exceed .70 if the scale has more than 14 items. Even higher values of α will be achieved if the dimensions are correlated. Finally, α may differ for measures collected during different periods of a longitudinal study. Both the variance in the true scores and the measured scores may change over time so that α can change dramatically. A measure of IQ collected on a group of children at age 4 will typically have a lower α than the same measure collected on the children at age 10. In later sections, we describe alternative approaches that address several of these issues as well as others that arise in longitudinal measurement contexts.

Test–retest correlations. A second method of estimating reliability is to calculate the correlation between the scores on the same set of items taken at two points in time. Test–retest approaches assume that (a) the participants' true scores do not change on the measure during the (short) interval between Time 1 and Time 2 and that (b) responding to the item at Time 1 has no effect on the response at Time 2 (e.g., no memory for prior responses on an ability test). Green (2003) has recently developed a test–retest version of α. Test–retest α eliminates sources of error that change across measurement occasions (e.g., daily mood changes), but otherwise shares the assumptions and properties of traditional α described earlier.

In longer-term studies, the interpretation of the test–retest correlation changes. It can no longer be assumed that there has been no change in the participants' true scores or that all participants change at the same rate. Children and adults change over time in their abilities, personality traits, and physical characteristics such as height and weight. In this case the test–retest correlation is an estimate of the stability of the measure—the extent to which the (rank) order of the participants at Time 1 is the same as the order of the participants at Time 2. Otherwise stated, the level of the measure (e.g., height) may change over time, but stability is shown to the degree that participants' amount of change is proportional to their initial level on the measure.

Homogeneity (Unidimensionality)

Interpretation of measures is greatly simplified if the measure assesses a single dimension (underlying factor). For example, imagine that a measure of college aptitude were developed. Unbeknownst to the test developers the items reflect a major dimension of IQ and a secondary dimension of conscientiousness. These two dimensions have only a minimal correlation. Both dimensions may predict good performance in many classes. But the conscientiousness dimension may be a far better predictor of performance in a history course in which large amounts of material must be regularly learned. In contrast, IQ may be a far better predictor of performance in a calculus course. By separating the two dimensions, we can gain a far greater understanding of the influence of the two dimensions in performance in different college classes. Indeed, the interpretation of the body of research associated with several classic measures of personality has been difficult because of the existence of multiple dimensions underlying the personality scale (see Briggs & Cheek, 1986; Carver, 1989; Neuberg, Judice, & West, 1997 for discussions). Finch and West (1997) discussed testing of measures in cross-sectional studies that are hypothesized to have more complex, multidimensional structures.

In longitudinal research, these issues only become more difficult because dimensions within a scale may change at different rates. For example, Khoo, Butner, and Ialongo (2004) found that a preventive intervention led to a linear decrease on a dimension of general aggression, but no change on a secondary dimension of indirect aggression toward property during the elementary school years. Such findings make it necessary to consider a more complex measurement structure in assessing longitudinal effects on the aggression scale.

The most commonly used method of assessing the dimensionality of measures in cross-sectional studies is confirmatory factor analysis (see Eid, Lischetzke, & Nussbeck, chap. 20, this volume; Hattie, 1985 for a review). In this approach, the researcher hypothesizes that a specific measurement model consisting of one or more latent factors underlies a set of items. The measurement model is then tested against data with two aspects of the results of the test being of special interest. (a) The procedure provides an overall χ^2 test (likelihood ratio test) of whether the hypothesized model fits the observed covariances between the items. If the value of the obtained χ^2 is *not* significant, then the hypothesized model fits the data. For large samples, the χ^2 test may reject even close-fitting models so that various fit indices such as the RMSEA and the CFI, which are less dependent on sample size, may be used to assess whether the model is adequate. (b) The strength of the relationship between the factor and each item (λ = factor loading) is estimated. In some models, the λs can be expressed in standardized form, in which case they represent the correlation between the latent factor and each item. Alternatively, one of the items may be treated as a reference variable ($\lambda = 1$). The strength of each of the other loadings is interpreted relative to the reference variable, values of λ >1 indicate a relatively larger change, and values of $\lambda < 1$ indicate a relatively smaller change in the measured variable corresponding to a one-unit change in the latent factor (see Steiger, 2002).

Confirmatory factor analysis can also be used to estimate coefficient alpha. We noted earlier that coefficient alpha assumes that all measures are equally good measures of the underlying construct. This assumption means that the factor loadings of all the items on the factor are equal, known as the assumption of essential tau equivalence. Comparing the fit of a model in which the λs are constrained to be equal, versus an alternative model in which the λs are freely estimated, tests essential tau equivalence. If the fit of the two models does not differ, then the assumption of essential tau equivalence is reasonable. McDonald (1999) and Raykov (1997) provide procedures for estimating α both when the assumption of essential tau equivalence is and is not met. Later in this chapter we will extend the idea of testing of assumptions about measurement structure to longitudinal data. To the extent measures have the same structure at two (or more) time points, the results of analyses using the measures become more interpretable.

Scaling

Stevens (1951) proposed an influential classification of measurement scales. Beginning with the lowest level in the hierarchy, nominal scales assign each participant to an unordered category (e.g., marital status: single, married, divorced, widowed). Ordinal scales assign each participant to one of several ordered categories (e.g., clothing size: 1 = small, 2 = moderate, 3 = large). Interval scales assign participants a number such that a one-unit difference at any point on the scale represents an identical amount of change (e.g., a change from 3 to 4 degrees or from 30 to 31 degrees represents the same change in temperature on the Celsius scale). Finally, ratio scales share the same equal interval property as the interval scale, but in addition have a true 0 point where 0 represents absence of the measured quantity (e.g., height in centimeters).

Stevens originally argued that the level of measurement limits the type of statistical analysis that may be performed. This position is potentially disturbing because many measures in psychology may not greatly exceed an ordinal level of measurement. Indeed, Krosnick and Fabrigar (in press) have shown that labels used to represent points on Likert-type items often do not come close to approximating equal spacing on an underlying dimension. On the other hand, several authors (e.g., Cliff, 1993; McDonald, 1999) have noted that for *t*-tests and analysis of variance, whether the measurement scale is ordinal, interval, or ratio, makes only a modest difference in the conclusions about the existence of differences between groups, so long as the assumptions of the analysis (e.g., normality and equal variance of residuals) are met. Similarly, for linear regression analysis or structural equation modeling, the level of measurement also does not have a profound effect on tests of the significance of coefficients. These results occur because monotonic (order preserving) transformations typically maintain a high correlation between scores on the original and transformed scales. Often, ordinal measurement will be "good enough" to provide an adequate test of the existence of a relationship or group difference even with statistical tests originally designed for interval level data.

However, if we have hypotheses about the *form* of the relationship between one or more independent variables and the dependent variable, ordinal measurement is no longer "good enough." Longitudinal analyses testing trend over time require interval level measurement. The origin and units of the scale must be constant over time; otherwise, the test of the form of the relationship will be confounded with possible effects of the measuring instrument. When standard statistical procedures designed for interval-level data are used with ordinal-level data, estimates of parameters of the growth model will be seriously biased. Special methods designed explicitly for ordinal-level data and large sample sizes are required (Mehta, Neale, & Flay, 2004).

Changes in the origin or units of the scale can happen because raters explicitly or implicitly make normative judgments relative to the participant's age and gender.[1] Consider the trait physically active. Informants may rate the second author as being very physically active—a rating of 8 on a 9-point scale ranging from "not at all" to "extremely" active at age 25 and then again at age 50. Yet, physical measures of activity (e.g., a pedometer) may show twice as much physical activity at age 25 as at 50. In effect, such ratings may be "rubber rulers" that correctly describe the standing of the individual *relative* to a same age comparison group. However, when changes occur in either the origin or the units of the scale, clear interpretation of the results of longitudinal analyses focused on the form of change is precluded. These problems do not characterize all longitudinal studies. Physical measures (e.g., height, blood pressure) and many cognitive measures provide invariant measurement at the interval level. Some rating scale measures may approximate interval-level measurement and be suitable for short-term longitudinal studies. But, few investigators consider this fundamental issue—the origin and units of the measure must be constant over time. Such invariance is fundamental in interpreting the results of longitudinal studies of change. We revisit this issue later in the chapter.

THREE LONGITUDINAL MODELS

At this point it would be beneficial to introduce several of the most common new longitudinal models for analyzing stability and change using continuous latent variables. These models include autoregressive models, trait–state models, and growth curve models.

Examining Stability: Autoregressive Models

Autoregressive models are used to examine the stability of the relative standing of individuals over time. Figure 21.1 illustrates an autoregressive model for a three-wave data set. In this data set (Biesanz, West, & Millevoi, 2004), 188 college students were assessed at weekly intervals on a measure of the personality trait of conscientiousness (Saucier & Ostendorf, 1999). According to Saucier and Ostendorf, conscientiousness is comprised of four closely related facets: orderliness, decisiveness, reliability, and industriousness. At each time period, we estimated the latent construct of conscientiousness. In the model presented in Figure 21.1, the factor loading of each facet was constrained to be equal over time so that the units of the latent construct would be the same at each measurement wave. Orderliness serves as the marker variable for the construct ($\lambda = 1$). λs for the other facets range from .62 to .67.

In the basic autoregressive model, the scores on the factor at Time t only affect the scores on the factor at Time $t + 1$. If there is perfect stability in the rank order of the students on the factor from one time period to the next, then the correlation will be 1.0, whereas if there is no stability, then the correlation will be 0. In the present example, there is considerable stability in the conscientiousness factor: the unstandardized regression coefficients are .78 (correlation = .85) for Week 1 to Week 2 and .84 (correlation = .88) for Week 2 to Week 3. These stabilities greatly exceed the corresponding simple test–retest correlations of .63 and .65, respectively.

Multiindicator autoregressive models have two distinct advantages over simple test–retest correlations. First, the model partitions the variance associated with the four indicators (facets) at each time into variance associated with the factor of conscientiousness and residual variance so that the stability coefficients are not attenuated by measurement error. Second, part of the residual variance may be due to a systematic feature of the facet (uniqueness) that is not shared with the latent construct of conscientiousness. Correlating the uniquenesses over

[1] For example, Goldberg's (1992) measure of the Big Five personality traits explicitly instructs informants to rate the participant relative to others of the same age and gender.

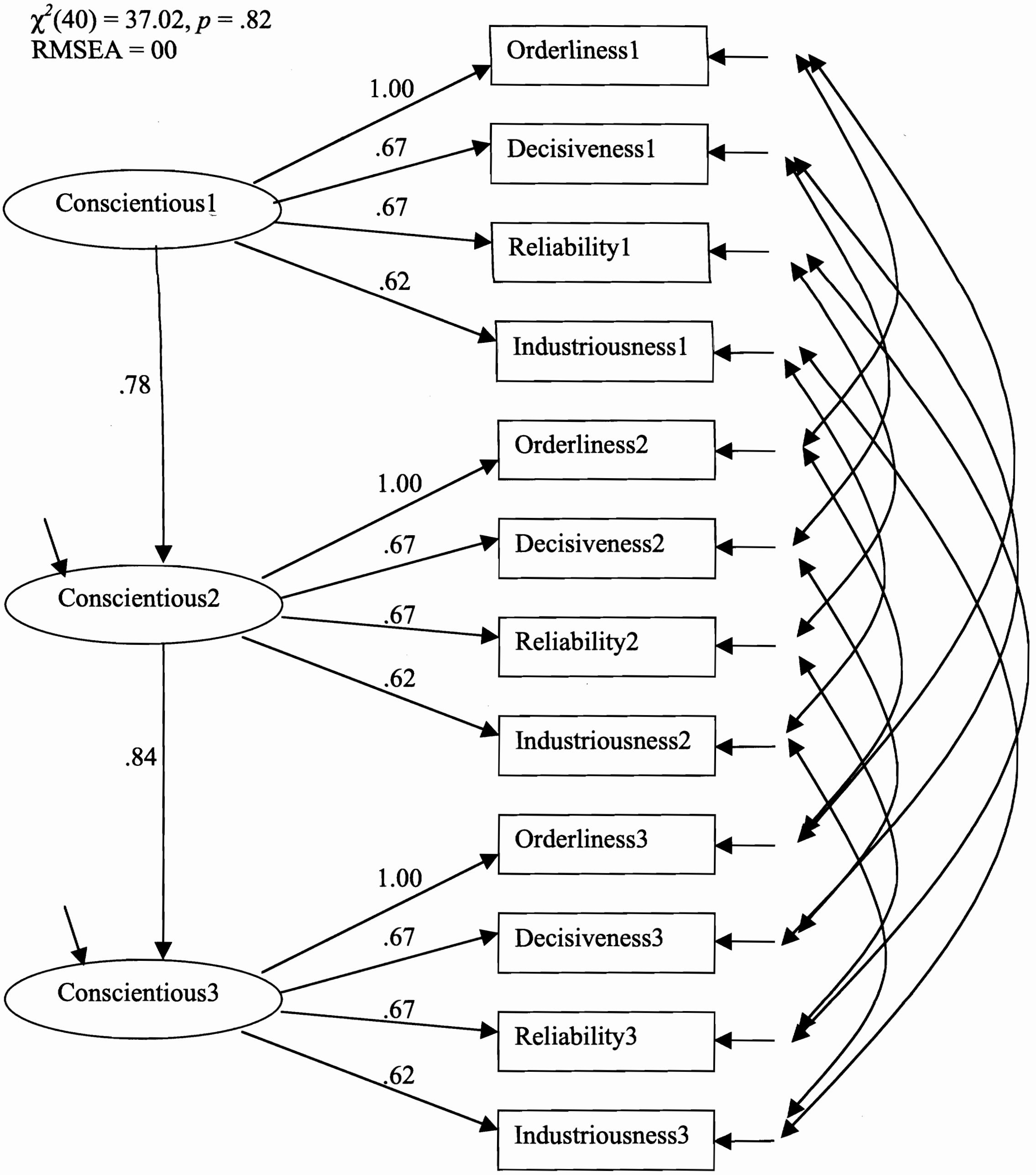

FIGURE 21.1. Autoregressive model.

each pair of time periods removes any influence of the stability of these systematic components of the residual. Otherwise, the estimate of the stability for the conscientiousness factor would be confounded by these unique components associated with each of the facets.

We estimated three alternative models to illustrate features of the model depicted in Figure 21.1. First, we investigated the effect of correlating the uniquenesses. Model (a), which included the correlated uniquenesses, showed a substantially better fit to the data, $\chi^2(40) = 35.1$, *ns*, RMSEA = .00, than Model (b), in which the correlations between the uniquenesses are deleted, $\chi^2(52) = 500.2$, $p < .0001$, RMSEA = .22). An RMSEA of .05 or less is typically taken as evidence of a close-fitting model. This result indicates that the correlated uniquenesses need to be included in the model. Second, we investigated the effect of constraining the factor loadings to be constant over time. Model (c), which is portrayed in Figure 21.1, also resulted in an acceptable fit to the data, $\chi^2(46) = 37.0$, ns, RMSEA = .00. The difference in fit between Models (a) and (c) may be directly compared based on their respective χ^2 and *df* values using the likelihood ratio test (Bentler & Bonett, 1980), $\chi^2(6) = 1.9$, ns. Given that the fit of the two models to the data does not differ, Model (c) is preferred both because it has fewer parameters (parsimony) and more importantly, because it simplifies interpretation by guaranteeing that the conscientiousness construct has the same units at each measurement wave.

Cross-lagged autoregressive models may be used to investigate the ability of one longitudinal series to predict another series. For example, Aneshensel, Frerichs, and Huba (1984) measured several indicators of illness and several indicators of depression every 4 months. The two constructs were modeled as latent factors. Moderate stabilities were found for both the illness and depression constructs. The level of depression at Wave *t* consistently predicted the level of illness at Wave *t* + 1, over and above the level of illness at the Wave *t*. In a similar study, Finch (1998) found that social undermining consistently predicted negative affect 1 week later over and above the level of negative affect the previous week. Such lagged effects show both association and temporal precedence, providing support for hypothesized direction of the causal relationship between the two variables (e.g., depression → physical illness). Jöreskog (1979) and Dwyer (1983) presented several useful variants of the basic autoregressive model for longitudinal data. Of importance, clear interpretation of the findings of these models assumes there is *not* systematic change in the level of the series of *measures (growth or decline) for each individual* over time (Willett, 1988). Curran and Bollen (2001) and McArdle (2001) have proposed models that combine growth and autoregressive components to address this issue.

Trait–State Models

Many important psychological phenomena (e.g., moods) appear to be influenced both by an individual's chronic level (trait) as well as temporary fluctuations from that chronic level (state). Latent trait–state models (Steyer, Ferring, & Schmitt, 1992; Steyer, Schmitt, & Eid, 1999; see Figure 21.2) partition each measure collected at each measurement occasion into three components. First is a component that represents the trait construct measured at a specific time point (denoted Time 1, Time 2, and Time 3 in Figure 21.2). This component is further partitioned into (a) a latent trait factor that characterizes the person's stable general level on the construct of conscientiousness (denoted as Consci in Figure 21.2) and (b) a latent state residual that characterizes temporary (state) effects on the person associated with each measurement wave. Second, the method factor represents the stable influence of the specific measure (here, the measure of each facet of conscientiousness, denoted Order, Decis, Reliab, Indust, respectively). Third, as in previous models, another component reflects random measurement error.

The latent state–trait model shows a good fit to the conscientiousness data, $\chi^2(39) = 31.87$, *ns*, RMSEA = .00). The clear partitioning of the observed scores on the measure into trait, state, measure, and error variance components provides a strong basis for predicting external criteria. For example, the relatively pure measure of the trait of conscientiousness that is estimated can be used to predict conscientiousness-related behaviors such as class attendance or worker productivity. The latent trait–state model can also partition the total amount of variance in the observed scores into trait, state, measurement method, and error variance components (see Steyer et al., 1992). In the present example, 42% of the variance in the observed scores is *associated with the stable latent trait factor for conscientiousness.*[2] Or, if the researcher were interested in situational effects on conscientiousness (e.g., if midterm exams were given prior to the Week 2 measurement), the proportion of the total variance

[2] The instructions emphasized answering based only on the past week's behaviors.

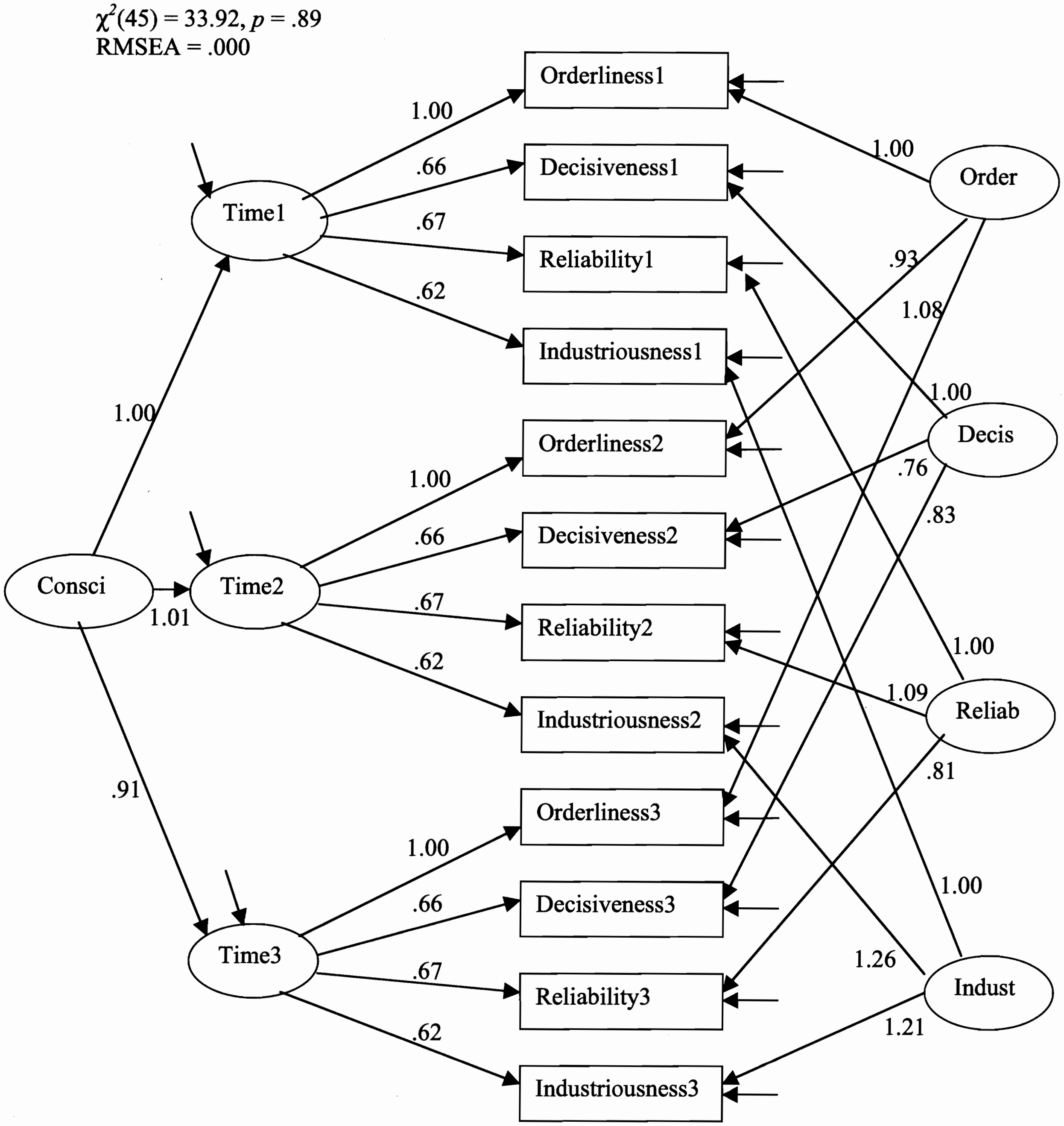

FIGURE 21.2. Latent state–trait model. Consci is the conscientiousness latent construct; Order, Decis (decisiveness), Reliab (reliability), Indust (industriousness) represent the four specific facets of conscientiousness.

in the observed scores associated with the latent state residuals could be computed. Steyer et al. (1992) discussed a variety of potential methods of partitioning the variance to produce estimates of several diverse forms of reliability and stability that may be useful in different research contexts. Steyer et al. (1992) and Kenny and Zautra (2001) compared several variants of the latent trait–state model.

Although the basic latent trait–state model has several important strengths, it also has three limitations. First, like the autoregressive model, the basic state–trait model focuses only on the relative ordering of a set of individuals. Clear interpretation of findings requires there is *not* systematic growth or decline for each individual over time. Otherwise, more complex models that combine growth and trait–state components are required (Tisak & Tisak, 2002). Second; the temporal ordering of the observations is not represented in the analysis. Otherwise stated, the data from any two time periods (e.g., 2 and 3) can be exchanged without affecting the fit or any important features of the model. Third, like multitrait–multimethod models (Eid, 2000; Kenny & Kashy, 1992), latent trait–state models can be difficult to fit with many data sets. Data sets with small state components or small method components can lead to improper solutions. In general, adding more time periods, more measures, and more participants appears to improve estimation. Steyer et al. (1999) present approaches that may be used when there are problems in estimation.

Growth Curve Modeling

In longitudinal studies with three or more measurement waves, growth curve modeling can provide an understanding of individual change (Laird & Ware, 1982; McArdle & Nesselroade, 2003; Muthén & Khoo, 1998). Researchers may study individual growth trajectories and relate variations in the growth trajectories to covariates that vary between individuals. They may also get better estimates of true growth by studying the effects of covariates that vary over time within individuals. We use the hierarchical modeling framework here to describe the models.

Conceptually, growth curve modeling has two levels denoted as Level 1 (within individuals) and Level 2 (between individuals). At Level 1 we describe each individual's growth using a regression equation. We focus here on the simplest model, linear growth. With linear growth we express the measure Y_{ti} of an individual i at time t as the sum of the individual's linear growth plus a residual ε_{ti} that represents random error at occasion t,

$$Y_{ti} = \alpha_i + \beta_i x_{ti} + \varepsilon_{ti} \quad ,t = 1,2,...,T \qquad (1)$$

In Equation (1), x_{ti} is the time-related variable such as age, measurement wave, or the elapsed time following the occurrence of an event (e.g., surgery). Note that x_{ti} has two subscripts, t and i, indicating it varies both over measurement occasions and across individuals. The intercept α_i represents the predicted level of Individual i on the measure when $x_{ti}= 0$. When time is scaled so that the first measurement occasion equals 0, α_i may be interpreted as the individual "initial status" or level on Y at the beginning of the study. The slope β_i represents the individual growth rate, the change in Y per unit of time. The individual intercept α_i and the individual slope β_i form a pair of growth parameters that characterize the individual trajectory. Figure 21.3 shows hypothetical linear growth curves of three individuals on a variable Y over time. Note that the individuals start at different levels (different α_is) and grow at different rates (different β_is). Other time-varying covariates may be added as predictors to the Level 1 equation.

For example, suppose we collected daily measures of stressful events w_{ti} and well-being Y_{ti} in each patient for 10 days immediately following minor surgery. We can add the time-varying covariate w_{ti} to Equation (1). For patient i, we now have

$$Y_{ti} = \alpha_i + \beta_i x_{ti} + \pi_i w_{ti} + \varepsilon_{ti} \quad ,t = 1,2,...,10 \qquad (2)$$

α_i is patient i's predicted well-being (initial status) at the completion of surgery; β_i is the rate of increase in well-being (slope). These parameters characterize each individual's growth function over and above the temporal disturbances accounted for by the time-varying covariate w_{ti}. π_i is the individually varying partial regression coefficient relating stress to well-being for Individual i, and ε_{ti} is the residual. Thus, Level 1 describes the change within individuals.

In the simplest Level 2 model, we assume that the set of α_is and the set of β_is are normally distributed. The means and variances of these growth parameters are estimated at Level 2. The means of the growth parameters allow us to obtain a mean trajectory for the whole group. To the extent that the variances of the growth parameters are greater than 0, there are differences between individuals in

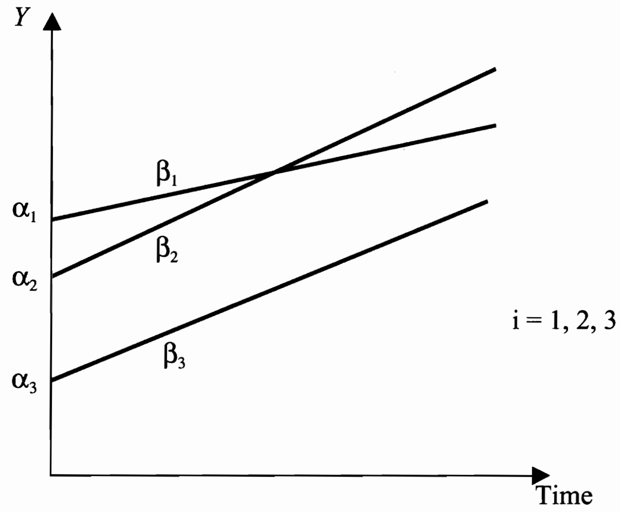

FIGURE 21.3. Growth trajectories for three individuals.

the growth patterns over time. With variation across individuals, the two individual growth parameters, α_i and β_i, can become outcome variables to be regressed on *time-invariant* individual background covariate variables. These background variables can be experimental treatment conditions (e.g., presurgical psychological intervention versus no intervention) or stable individual difference variables (e.g., neuroticism). The Level 2 equations for the intercepts and the slopes may be expressed as

$$\begin{aligned} \alpha_i &= \alpha_0 + \gamma_\alpha Z_i + \delta_{\alpha i} \\ \beta_i &= \beta_0 + \gamma_\beta Z_i + \delta_{\beta i} \end{aligned} \qquad (3)$$

where α_0 is the grand intercept (mean intercept across *N* individuals), β_0 is the grand slope (mean slope across *N* individuals), and Z_i is the time-invariant covariate (e.g., neuroticism) and $\delta_{\alpha i}$ and d_{bi} are the residuals associated with α_i and β_i respectively; and γ_α and γ_β are the regression coefficients. Besides the linear growth parameters, additional Level 2 equations may be written to account for variation in the Level 1 regression coefficients for the time-varying variables (e.g., daily stress) if these are included in the model. Thus, at Level 2, we model between individual differences in the values of the growth parameters (intercept and slope) and the regression coefficients for the time-varying variables.

Although we have focused on linear growth, more complex patterns including quadratic growth, growth to an asymptote, and other nonlinear forms of growth may be modeled as the number of measurement waves increases (Cudeck, 1996; Singer & Willett, 2002). In addition, different time-related metrics may be of focal interest such as age or elapsed time since an event (e.g., surgery) or the beginning of a developmental period (see Biesanz et al., 2003).

Standard growth curve models can also be estimated using structural equation modeling (Muthén & Khoo, 1998; Willett & Sayer, 1994). Mehta and West (2000) noted that the two approaches can both typically be used and produce the same results, but that some applications may be more amenable to one of the approaches. The hierarchical modeling approach discussed in this section may be more flexible in representing some nonlinear forms of growth. In contrast, the structural equation modeling approach often has more flexibility in modeling the measurement structure using multiple indicators of a construct at each time point and in modeling complex relationships between multiple series. Within the structural equation approach, features of autoregressive models (Curran & Bollen, 2001; McArdle, 2001) and features of latent trait–state models (Tisak & Tisak, 2002) can be combined with growth models.

The modeling of change using growth curve modeling described earlier calls for several very strong assumptions regarding the measurement scale. First, the repeated measurements must be made on at least an interval-level scale. Otherwise, the form of growth will be confounded by changes in the size of the measurement unit at each point in the scale. Second, there must be measurement invariance over time—the relationship between the observed measures and the underlying construct must remain constant with the passage of time. For example, items such as pushing and biting might measure physical aggression at age 4. However, at age 16 these items will no longer adequately reflect aggression, precluding meaningful study of change over time. On the other hand, if we measure aggression at age 16 with items like "threaten with gun or knife" and "hit with objects," then the meaning of the construct has changed. (See Patterson, 1995, on developmental change in constructs.) In such cases in which the items on instruments do change over the course of the study (e.g., different items on a measure of math ability in first and fourth grades), there is a need to ensure that the meaning of the construct remains the same. Educational researchers have been successful to some extent in the area of assessing skills and knowledge using vertical equating of overlapping test forms of increasing difficulty levels (see section on vertical equating). Similar techniques are not as well developed for longitudinal studies of psychological and affective constructs.

Other Longitudinal Models

Our emphasis has been on several of the more common new longitudinal models for stability and change using continuous latent variables. New models for other forms of data have also been developed. Space considerations did not permit us to consider longitudinal modeling of discrete latent classes (Langeheine, 1994; Lanza, Flaherty, & Collins, 2003), combinations of continuous and discrete latent variables (Muthén, in press), longitudinal models for single subjects (Browne & Zhang, in press; West & Hepworth, 1991), or the linear logistic model with relaxed assumptions for measuring change (Fischer & Formann, 1982).

MEASUREMENT OF CHANGE

For researchers who are interested in quantitative change over time rather than (rank order) stability, the measurements need to be made on a common scale that achieves at least an interval level of measurement over time. This property characterizes many physical measurements such as height, blood pressure, or counts of behaviors. However, this property often does not characterize psychological measures of attitudes and traits. Attempts to measure abilities, attitudes, or traits usually rely on the collective strength of responses to individual items within instruments. In the measurement of psychological traits, the response to each item is typically assessed by either using a dichotomous response (e.g., "I enjoy parties"—true or false) or a Likert-type response scale that is essentially ordinal (e.g., "How much do you like parties?" rated on a 5-point scale from "not at all" to "very much"). In current research practice the same instrument is administered at each measurement wave, and the total scale score at each wave is used to model change. However, this practice involves several important untested assumptions: (a) the scale is unidimensional, (b) the total scores yields interval level measurements, (c) the same total score would indicate the same construct level over time, and (d) there is measurement invariance over time. These assumptions are seldom checked or addressed.

If the measurements are not made on an interval scale, equal differences in scores over time at different levels of the construct may *not* mean the same amount of change in the construct. The measurement unit stretches or shrinks as a function of the level that is measured—the rubber ruler problem. Desirable interval scale properties can usually be achieved through careful scale construction and through successfully applying measurement models.

External Scale Construction: Rasch and Item Response Theory Modeling

Several methods exist for developing strong measurement scales separately from the longitudinal model of stability or change (see Rost & Walter, chap. 18, this volume). These methods can be applied to dichotomous or ordinal data. The scales can be developed using the same or a different data set from that used to test the longitudinal model. The Rasch model (1-parameter; Rasch, 1960; Wright & Masters, 1982; Wright & Stone, 1979) provides interval-level measurement, and the 2-parameter logistic Item Response Theory model (IRT; see Embretson & Reise, 2000) provides a good approximation to interval-level measurement when the data are consistent with the model. These are probabilistic measurement models. For dichotomous items, equal changes in the underlying latent construct correspond to equal changes in the log of the odds of endorsing an item, for any level of the latent trait.

For items with multiple ordered response categories (1 = "not at all," 2, 3, 4, 5 = "very much) that typify Likert-type scales, there are extensions of both the Rasch and the 2-parameter IRT models.

A variety of polytomous models for multiple-ordered response categories have been developed. The Rasch extensions include the partial credit model (Masters, 1982) and the rating scale model (Andrich, 1978). The 2-parameter IRT extensions include the graded response model (Samejima, 1969) and the modified graded response model (Muraki, 1990). The basics of the Rasch model and its extensions are described and illustrated by Rost and Walter (chap. 18, this volume). Drasgow and Chuah (chap. 7, this volume) explain and illustrate the 2- and 3-parameter models in detail. In each of these models, there are multiple probability curves for each item, one for each response category. These probabilities provide information on how each category functions relative to other categories within an item. These models produce good approximations of interval level score estimates of the underlying construct while treating the response categories as ordinal. The interval level score estimates produced can be used to model longitudinal change.

Simultaneous Longitudinal and Measurement Modeling

Structural equation modeling permits simultaneous modeling of the measurement structure and the longitudinal model of stability or change. In the measurement portion of the model, each latent construct is hypothesized to be error free and normally distributed on an interval scale. The structural part represents the relationships between the latent constructs. This modeling approach can also be extended to two or more ordered categories (Muthén, 1984). This approach assumes that each dichotomous or ordered categorical measured variable is characterized by an underlying normally distributed continuous variable. For each measured variable, *c*-1 thresholds are estimated that separate each of the *c* categories (e.g., one threshold for a dichotomous variable). If the assumptions are met, then Muthén's approach will provide estimates of the underlying factors that approximate an interval-level scale of measurement. Indeed, Takane and de Leeuw (1987) have shown that 2-parameter IRT models and confirmatory factor models are identical for dichotomous items under certain conditions. Unfortunately, large sample sizes (e.g., 500–1,000 or more cases) are often required for the appropriate use of structural equation modeling approach to categorical data. Newer estimation methods may offer promise of adequate estimation with smaller sample sizes (Muthén & Muthén, 2004). However, separate scale development using external methods such as Rasch or IRT modeling will often be more efficient.

MEASUREMENT INVARIANCE ACROSS TIME

In cross-sectional research, a major concern is addressing the issue of measurement invariance across groups. Does a set of items measure cognitive ability equally well in African-American and Caucasian populations? Does a standard measure of extroversion or depression capture the same underlying construct in the United States and China? Similar issues can arise in longitudinal research when measures are collected over extended periods of time. Does a standard measure of childhood extroversion assess the same construct at age 12 and age 18? If change over time is to be studied, the same construct must be measured at each time point. Measurement invariance may be established within either (a) the Rasch/IRT or (b) the confirmatory factor analysis approaches.

Measurement invariance implies that the score on the instrument is independent of any variables other than the person's value on the theoretical construct of interest. To illustrate how measurement invariance might fail, consider a test of mathematics ability for intermediate school students. Suppose that the following item were devised: "A baseball player has 333 at bats and 111 hits. What is his batting average?" Although this item clearly reflects mathematical ability, it also reflects knowledge about baseball—knowledge that is more likely to be found in male than female students with the same level of mathematics ability. Such items that exhibit a systematic relationship with group characteristics after controlling for the construct level are said to be functioning differentially across groups. Differential item functioning (DIF) thus contributes to measurement *non*-invariance across groups. Similarly, if measurement invariance holds across time, then the probability of a set of observed scores occurring is conditional only on the level of the latent construct and is independent of any variable related to time:

$$P(\mathbf{Y} \mid \theta, \mathbf{X}_t) = P(\mathbf{Y} \mid \theta),$$

where Y is the set of observed scores, θ is the level of latent construct and X_t is the set of time-related variables such as age and testing occasion. For example, an item such as, "Did you make your bed this morning?" might be a good measure of the orderliness facet of conscientiousness for college students at the beginning of the semester, but not during exam weeks. Only when measurement invariance over time is established can we conclude that the measurement scale for the underlying construct remains the same. Of importance, measurement invariance allows us to conclude that changes in scores are the result of changes over time on the construct of interest rather than on other characteristics of the instrument or the participants.

Rasch and IRT Approaches

For unidimensional constructs with dichotomous or ordered categorical items, the Rasch model and the 2-parameter logistic IRT model are commonly used (see Embretson & Reise, 2000). The Rasch model has one parameter (b_j) for each item representing its difficulty (level), whereas the two-parameter IRT model has both a difficulty parameter (b_j) and a discrimination (slope) parameter (a_j) for each item. Assessment of measurement invariance across time involves checking that the item parameters a_j and b_j have not changed over time. If the data fit the Rasch model, $a_j = 1$ for each item so only the set of b_js will be checked. For measures with multiple ordered categories, the item parameters corresponding to each possible response category will need to be checked for each item. These procedures work very well for unidimensional scales that are often developed for the assessment of abilities. Unfortunately, current measures of many psychological constructs (e.g., many attitudes; traits) are very often multidimensional, consisting of several underlying factors or a major factor and several minor factors. The use of Rasch and IRT procedures for the assessment of measurement invariance is not as well studied for multidimensional psychological scales.

Confirmatory Factor Analysis Approaches

When data are continuous and there are one or more underlying factors, confirmatory factor analysis procedures may be used to test measurement invariance. Meredith (1993) considered the issue of measurement invariance across groups, and he developed a sequence of model comparisons that provide a close parallel to the IRT approach. Widaman and Reise (1997) presented a clear description of these procedures, and Meredith and Horn (2001) have recently extended this approach to testing measurement invariance over time. In brief, a hierarchical set of models with increasingly strict constraints are compared. First, a baseline model is estimated. In this model, the value of the factor loadings of each measured variable on an underlying construct may differ over time. For example, consider the model of conscientiousness (Figure 21.1) discussed in the prior section on "examining stability: autoregressive models." Suppose we had allowed the factor loadings to vary over time (Model 1) and this model fit the data. Such a model, known as a configural model, would suggest that similar constructs were measured at each measurement wave. In contrast, imagine that although the single factor of conscientiousness fit the data adequately at Wave 1, over the course of a longer-term study the conscientiousness factor split into two separate factors—one factor representing orderliness and reliability and a second factor representing decisiveness and industriousness. Such a result would indicate the fundamental nature of the conscientiousness factor had changed over time (failure of configural invariance), making difficult any interpretation of stability or change in conscientiousness.

When the configural model fits the data (as in our earlier example), we can investigate questions related to the rank-order stability of the general construct. Note, however, that the conscientiousness latent construct (factor) at each measurement wave would not necessarily be characterized by a scale with the same units. To establish that the units are identical over time, we need to show that the factor loadings are equal across time. As we saw in the model represented in Figure 21.1, the imposition of equal factor loadings did not significantly affect the fit of the model in our example. Thus, our study of stability was improved by our ability to correlate constructs measured using the same units at each measurement wave.

Finally, suppose that we wish to establish that the scale of the construct has both the same units and the same origin over time (i.e., interval level of measurement). Recall that this condition must be met for proper growth modeling. To illustrate differences in the origin, consider that the Celsius and Kelvin temperature scales have identical units (1 degree difference is identical on both scales). However, the origin (0 degrees) of the Celsius scale is the freezing point of water, whereas the origin of the Kelvin scale is absolute 0 (where molecular motion stops). To establish that the origins are identical, we need to consider the level of each measured variable (mean structure) in addition to the covariance structure. If the origin of the scale does not change over time, then the intercept (the predicted value on each measured variable when the level of the underlying construct $\theta = 0$) also must not change over time. If the fit of a model in which the intercepts for each measured variable are allowed to vary over time does not significantly differ from that of a more restricted model in which the each variable's intercept is constrained to be equal over time, this condition is established.[3] If this condition can be met, then the level of measurement invariance over time necessary for proper growth curve modeling has been established. Widaman and Reise (1997) discussed still more restrictive forms of measurement invariance that can be useful in some specialized applications. Muthén (1996), Mehta et al. (2004), and Millsap and Tein (in press) present extensions of the confirmatory factor analysis approach that can be used to establish measurement invariance for multidimensional constructs measured by dichotomous or ordered categorical measures.

VERTICAL EQUATING: ADDRESSING AGE-RELATED CHANGES IN ITEM CONTENT

The items required to measure a latent construct can change as participants age. In educational research children are expected to acquire knowledge and learn appropriate skills. For example, in a test of mathematical proficiency, items related to multiplication may be needed in third grade, whereas items related to fractions may be needed in sixth grade. The test forms for each grade level must be equated onto a single common metric to measure educational progress. Vertical equating must be achieved externally prior to any longitudinal modeling of the data.

Vertical equating uses Rasch models or the 2-parameter IRT models to calibrate tests onto a single common "long" interval scale. This "long" scale covers the full range of proficiency as assessed using easier tests in the lower grade levels and more difficult tests in the higher grade levels. The equating of test forms is made possible by embedding common item sets in the test forms. The common item sets serve as "anchor" or "link" items for the equating. Any change in the probability of getting each item correct should only occur if there is a change in the individual's level on the underlying construct; otherwise, the item is showing DIF as a function of grade level. For example, an item that is assessing problem-solving skills at Grade 2 but is just assessing routine skills at Grade 4 may very likely show DIF. Even though the wording of the item is identical, this item functions differently across the two different grades and will not make a good link item. Thus, for unidimensional constructs vertical equating combines testing for DIF and establishing measurement invariance of link items and linking scales (see Embretson & Reise, 2000). Applications of these equating procedures permit the development of computerized adaptive tests (see Drasgow & Chuah, chap. 7, this volume) that select the set of items that most precisely assess each participant's level on the underlying latent construct θ. Unfortunately, vertical equating of multidimensional constructs is difficult to achieve because the rate or form of growth may vary across dimensions so that common item set(s) that adequately represent each of the dimensions cannot always be constructed.

[3] The full confirmatory factor analysis model including mean structure can be expressed as $Y = \nu + \Lambda\eta + \varepsilon$. Y is the p x 1 vector of observed scores, ν is p x 1 vector of intercepts, η is the m x 1 vector of latent variables, Λ is the p x m matrix of the loadings of the observed scores on the latent variables η, and ε is the p x 1 vector of residuals. For modeling longitudinal measurement, a model in which both Λ and ν are constrained to be equal over time must fit the data.

In contrast to research on measures of educational progress and abilities, far less attention has been given to equating psychological constructs like traits and attitudes across age. Typically, the same instrument is used at each measurement wave to assess individuals on a construct of interest. This practice is often appropriate when the time spanning the study is relatively short and the study does not cross different periods of development. If the reading level and the response format are appropriate for the participants over the duration of the study, serious age-related problems with the instrument are unlikely to occur. However, when a measure crosses developmental periods, for example, in a study that follows subjects from adolescence to young adulthood, the instrument may not capture the same construct adequately as subjects mature. Some items may need to be phased out over time while other items are being phased in. What results are instruments that are not identical, but that have overlapping items for different developmental periods. For example, the Achenbach Youth Self-Report externalizing scale (YSRE) was developed for youth up to age 18 (Achenbach & Edelbrock, 1987), and the Young Adult Self-Report externalizing scale (YASRE) was developed for young adults over age 18. Each measure has approximately 30 items, yet only 19 of these items are in common across the two forms. If participants were administered the two forms of the YSRE during a longitudinal study that crossed these developmental periods, the two forms would need to be equated onto a common scale if growth is to be studied. Such vertical equating of psychological measures is rare.

Many of the standard measures used in psychology were designed for cross-sectional studies to examine differences between individuals; they were not developed for the study of change within an individual across time. As an illustration, many traditional instruments used for research in developmental psychological are normed for the different ages. Norm-referenced metrics do *not* comprise an interval scale and are often not suitable for capturing change. One example of a norm-referenced metric is the grade-equivalent scale (e.g., reading at a fifth-grade level) used in measuring reading achievement. Seltzer, Frank, and Bryk (1994) compared growth models of reading achievement using the grade equivalent metric and using interval-level scores based on Rasch calibration. They found that the results were very sensitive to the metrics used.

Theoretically, structural equation modeling approaches could also be used for vertical equating. However, McArdle, Grimm, Hamagami, and Ferrer-Caja (2002) noted that such efforts to date with continuous measures have typically involved untestable assumptions and have often led to estimation difficulties. At the same time, studies to date have not carefully established common pools of items (or subscales) that could be used to link the different forms of the instrument. Mehta et al. (2004) addressed vertical equating of ordinal items.

CONCLUSION

Researchers have increasingly recognized the value of longitudinal designs for the study of stability and change, for understanding developmental processes, and for establishing the direction of hypothesized causal effects. Researchers have increasingly gone beyond the minimal two-wave longitudinal design and now often include several measurement waves. These multiwave designs potentially permit the researcher to move beyond traditional analyses such as correlation, regression, and analysis of variance and use promising newer analysis approaches such as the autoregressive, latent state–trait models, and growth curve models presented in this chapter.[4] These analyses can potentially provide better answers to traditional questions in longitudinal research. They also permit researchers to raise interesting new questions that were rarely, if ever, considered within the traditional analytic frameworks. For example, latent trait–state models can provide definitive information about the role of states and traits, a classic problem in personality measurement. Growth curve models permit researchers to identify variables that explain individual differences

[4] Ferrer and McArdle (2003) and McArdle and Nesselroade (2003) provide a review of these and several other recently developed longitudinal models that could not be included in this chapter because of space limitations.

in growth trajectories, a question that was not raised until the development of these models.

Longitudinal researchers, like researchers in many other areas of psychology (see Aiken, West, Sechrest, & Reno, 1990) have often paid minimal attention to measurement issues. And historically, such lack of attention could be justified because the traditional measurement practices were "good enough" to provide adequate tests of the hypotheses. Answering questions within a traditional null hypothesis testing framework about the simple existence of a difference between means or of a correlation does not require sophisticated measurement. Ordinal level measurement provides sufficient information. And statistical methods like ANOVA and regression that were designed for interval-level scales have proven to be relatively robust even when applied to ordinal scales. So long as the assumptions of the procedure (residuals are independent, normally distributed, and have constant variance) are met, the traditional measures produce reasonable answers (Cliff, 1993). And researchers could compensate for the loss of statistical power associated with the use of ordinal measurement by moderate increases in sample size. However, psychologists have begun to ask more complex questions about the size and the form of relationships. What is the magnitude of the effect of treatment? How much do boys versus girls gain in proficiency in mathematics achievement from Grade 1 to 3? Does the acquisition of vocabulary in children between 12 and 24 months show a linear or exponential increase? Proper answers to such questions require more sophisticated measurement.

There is an intimate relationship between theory, methodological design, statistical analysis, and measurement. Many traditional questions about the *stability* of constructs and the relationship of one construct to another over time can be adequately answered even without achieving interval-level measurement. Some added benefits do come from interval-level measurement: More powerful statistical tests and a more definitive interpretation of exactly what construct is or is not stable (and to what degree) can be achieved. But, in contrast, as psychologists ask increasingly more sophisticated theoretical questions about *change* over time and use more complex statistical analyses that are capable of providing answers to these questions, interval-level measurement will be required. The exemplary initial demonstrations of the newer statistical models for modeling change have deliberately used interval-level measures. To cite two examples, Cudeck (1996) reported nonlinear models of growth in physical measures (e.g., height) and number of correct responses in learning. McArdle and Nesselroade (2003) emphasized growth models using a Rasch-scaled cognitive measure (the Woodcock–Johnson measure of intelligence). As these newer statistical models of growth are applied to current measures of psychological characteristics (e.g., attitudes, traits), the limitations of many current measures will become more apparent. For example, how can researchers distinguish between linear growth and growth to an asymptote if they cannot be confident that measurements have been made on an interval scale? Evidence of measurement quality traditionally cited in reports of instrument development—adequate coefficient alpha, test–retest correlation, and correlations with external criteria—will not be sufficient for longitudinal researchers who wish to model growth using the newer statistical models that demand interval-level measurement.

In this chapter we have emphasized four features of longitudinal measurement for psychological characteristics. These features can be viewed as desiderata that can help ensure that the measurement of constructs over time is adequate for the study of growth and change. These desiderata can be achieved using Rasch or IRT approaches for dichotomous or ordered categorical items and confirmatory factor analysis procedures for continuous items.

1. Scales developed to measure the construct of interest should ideally be unidimensional. In cross-sectional studies, the use of scales with more than one underlying dimension has led to considerable complexity in the interpretation of the results of studies using these scales. Although multidimensional scales may be used in longitudinal studies, interpretation will be challenging because each of the underlying dimensions may change at different rates over time.

2. Scales should attempt to achieve an interval level of measurement. The same numerical difference at different points on the scale should indicate the same amount of change in the underlying construct.
3. Measurement invariance over time should be established to ensure that the construct has a stable meaning. Each of the items on the instrument should measure the same construct at each measurement wave. The goal is to produce measures that assess only change on the construct and not differential functioning of items as their meaning changes over time.
4. Measures should use items and response formats that are appropriate for the age or grade level of the participants. The different forms of the measure must be linked and equated onto a single common scale. This practice is commonly used in educational research where procedures for vertical equating of measures containing both different and overlapping items have been well developed. For psychological measures, this issue of externally developing age-appropriate measures will often arise in longer duration studies that cross different developmental periods.

Achieving these desiderata will provide a different degree of challenge for different areas of longitudinal research in psychology. Some existing areas such as the study of physical growth and the growth of cognitive abilities have long used measures that meet these desiderata. Emerging areas will need to ensure that they address these issues as they develop new measurement scales. And in many other existing areas researchers will need to rescale existing instruments to develop measures that more adequately meet these desiderata. But, in each case, there will be a clear payoff. Researchers will have a substantially enhanced ability to ask and properly answer interesting new questions about change in important psychological constructs.